Ira Narang, in2933

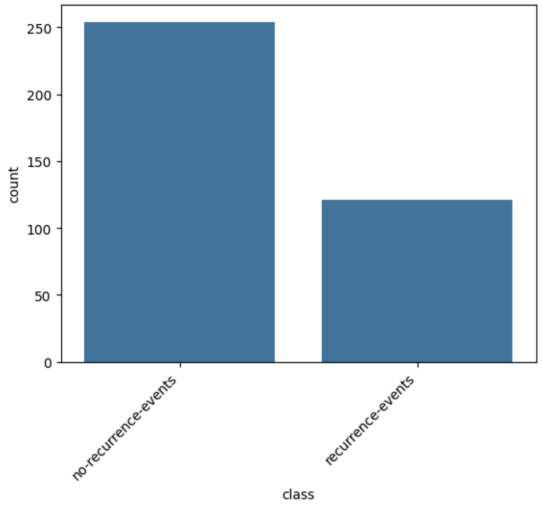
Project 1: Breast Cancer Classification

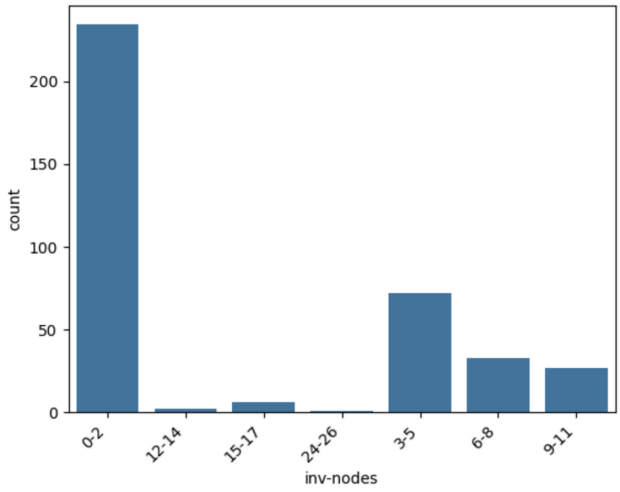
Data Preparation:

To prepare the dataset for classification, several preprocessing steps were performed. The dataset's shape and size were examined to understand its structure. Some categorical variables were converted to appropriate data types for analysis and modeling. Missing values in the dataset were treated using the mode within each age group to ensure consistency. Duplicate entries were identified and removed to avoid bias. Exploratory Data Analysis was conducted using visualizations such as count plots and box plots to identify trends and class distributions. Finally, categorical variables were transformed using one-hot encoding to ensure compatibility with machine learning algorithms.

Insights:

From the data preparation phase, several insights were gained. For example, most cases in the dataset had 0-2 affected lymph nodes, but some instances had significantly higher counts. The degree of malignancy distribution was slightly skewed, with a majority of cases exhibiting higher malignancy grades. The dataset showed an imbalance in the recurrence class distribution, where non-recurrence events were more frequent than recurrence events. These are only a few, but important, observations made from the generated plots. The plots for recurrence events and affected lymph nodes are shown below.





Training Procedure:

The dataset was split into training and test sets while maintaining class proportions to ensure fair model evaluation. Three classification models were trained: K-Nearest Neighbors (KNN) with n_neighbors=3, KNN with Grid Search CV to find the optimal k value, and a linear classification model for comparison with non-linear approaches. Each model's performance was evaluated using accuracy, recall, precision, and F1-score.

Model Performance:

K-Nearest Neighbor (KNN) Classifier Results:

	Accuracy	Recall	Precision	F1 score
Train	0.77	0.49	0.71	0.58
Test	0.62	0.28	0.37	0.32

The KNN model with k=3 showed good performance on the training set but suffered from overfitting, as evidenced by a drop in test accuracy (0.62) compared to training accuracy (0.77). Recall on the test set was 0.28, indicating that the model struggled to identify recurrence cases.

KNN Classifier using Grid search CV Results:

	Accuracy	Recall	Precision	F1 score
Train	0.69	0.12	0.67	0.20
Test	0.67	0.08	0.43	0.14

By applying Grid Search CV, the optimal k value (17) was found, leading to a model with improved test accuracy (0.67) but a lower recall score (0.08), suggesting a trade-off between overall accuracy and the ability to correctly identify recurrence cases.

Linear Classification:

	Accuracy	Recall	Precision	F1 score
Train	0.50	0.67	0.43	0.53
Test	0.61	0.61	0.34	0.44

The linear classifier provided a different perspective, achieving a test accuracy of 0.50 and a training accuracy of 0.61. Although its recall score on the test set was 0.61, its precision was low (0.34), resulting in an F1 score of 0.44. This suggests that the model was more likely to correctly identify recurrence cases but at the cost of an increased number of false positives.

Recall is crucial as it minimizes false negatives, ensuring recurrence cases are detected. This prevents delayed treatment which can be detrimental in this scenario. Despite lower accuracy, the linear model achieves the highest recall, making it the optimal choice, even with a higher false positive rate.

Model Confidence:

The final model demonstrated moderate predictive ability, particularly after applying Grid Search CV. However, given the nature of the problem, further improvements could be made by exploring additional models such as decision trees or ensemble methods. While the model provides valuable insights, real-world deployment should involve continuous validation of new data and collaboration with medical professionals to minimize risks.